UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 09-55
Economic Series (34)
July 2009

Departamento de Economía
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 916249875

"A COMPARISON OF THE SCIENTIFIC PERFORMANCE OF THE U.S. AND THE EUROPEAN UNION AT THE TURN OF THE XXI CENTURY"

Pedro Albarrán*, Joan Crespo**, Ignacio Ortuño*, Javier Ruiz-Castillo*
* Departamento de Economía, Universidad Carlos III
** Departamento de Economía Cuantitativa, Universidad Autónoma de Madrid

Abstract

In this paper, scientific performance is identified with the impact journal articles achieve through the citations they receive. The empirical exercise refers to 3.6 million articles published in 1998-2002 in 22 scientific fields, and the more than 47 million citations they receive in 1998-2007. The first finding is that a failure to exclude co-authorship among member countries within the EU (European Union) may lead to a serious upward bias in the assignment of articles to this geographical area. In the second place, standard indicators, such as normalized mean citation ratios, are silent about what takes place in different parts of the citation distribution. Consequently, this paper compares the publication shares of the U.S. and the EU at every percentile of the world citation distribution in each field. In 15 disciplines, as well as in all sciences as a whole, the EU share of total publications is greater than that of the U.S. one. But as soon as the citations received by these publications are taken into account the picture is completely reversed. The mean citation rate in the U.S. is greater than in the EU in every one of the 22 fields. In seven fields, the initial gap between the U.S. and the EU widens up as we advance towards the more cited articles, while in the remaining 15 fields –except for Agricultural Sciences– the U.S. always surpasses the EU when it counts, namely, at the upper tail of citation distributions. For all sciences as a whole, the U.S publication share becomes greater than that of the EU one for the top 50% of the most highly cited articles.

Acknowledgements

P. Albarrán and J. Ruiz-Castillo acknowledge financial support from the Spanish MEC, the first through grants SEJ2007-63098 and SEJ2006-05710, and the second through grant SEJ2007-67436. The database of Thomson Scientific (formerly Thomson-ISI; Institute for Scientific Information) has been acquired with funds from Santander Universities Global Division of *Banco Santander*. This paper is part of the SCIFI-GLOW Collaborative Project supported by the European Commission's Seventh Research Framework Programme, Contract number SSH7-CT-2008-217436.

I. INTRODUCTION

I. 1. Motivation

This paper compares the scientific performance of the U.S. and the EU, namely, the 15 countries forming the European Union before the 2004 accession. Like all other contributions to this literature referenced below, scientific performance is identified with the citation impact achieved by articles published in more than 8,000 academic or professional journals in 36 languages indexed by Thomson Scientific (TS hereafter). As far as data is concerned, the difference with other studies is that this paper uses a large sample consisting of almost 8 million articles published in 1998-2007, as well as the approximate 65 million citations received by them during that period. The articles belong to the 20 natural sciences and the two social sciences distinguished by TS.

This contribution is motivated by the idea that the design of a good science and technology policy for any area should start from an accurate diagnosis of the situation that, at least at some general level, can be shared by all agents involved in the chain that goes from the policy maker to the individual scientist, including experts in the evaluation of these activities. The problem is that it would appear that the issue of the relative scientific situation of the EU and the U.S. is not yet completely settled.

The goal launched at the often-quoted 2000 Lisbon meeting by the Council of the EU ("to become in 2010 the most competitive and dynamic knowledge-based economy in the world"), would seem to reveal an urge to change the European ways in the face of two worrisome circumstances: a sizable scientific gap with the U.S. dating at least from the middle of the last century, and the awakening of several developing countries in Asia that will surely become formidable rivals to everyone in some scientific and technological fields early in the XXI century. Against this view, the *First European Report on Science and Technology Indicators* (EC,

1994) popularized what became known as the "European Paradox", according to which Europe plays a leading world role in terms of scientific excellence but lacks the entrepreneurial capacity of the U.S. to transform it into innovation, growth, and jobs.[1]

If instead of consulting the official EU reports we turn to the short but interesting academic literature on the scientific wealth of nations, we also find rather different summary views about the first axes of the so-called European Paradox. For example, King (2004) indicates that *"The United States easily heads the list of nations in the volume of publications and citations and the share of top 1% cited papers, although the EU15 countries now publish more papers than the United States and are not far behind on citations"* (p. 311). Furthermore, *"…comparing citations between the United States and the EU15 shows that the gap between the two has shrunk significantly since May (1997)'s analysis based on figures up to 1993. The EU now matches the United States in the physical sciences, engineering and mathematics, although still lags in the life sciences"* (p. 316). Similarly, Shelton and Holdridge (2004) indicate: *"While the U.S. leads in most input indicators, output indicators may be more specific for determining present leadership. They show that the EU has taken the lead in important metrics and it is challenging the U.S. in others"* (Abstract, p. 353). These authors conclude *"So who is leading the world in Science and Technology: the U.S or the EU? While no single nation rivals the U.S. for the lead, it is becoming clear that the European Union as a whole is mounting a serious challenge."* (p. 362). On the other hand, Dossi *et al.* (2006) forcefully argue for the opposite view: *"The general conclusion from the bibliometric data is therefore far from supporting any claim to European leadership in science. On the contrary, one observes a structural lag in top-level science vis-a-vis the US, together with (i) a few sectoral outliers in physical sciences and engineering, and (ii) a few single institutional outliers (such as Cambridge in*

[1] In the same vein, see Chapter 4 –under the title "Beyond the European Paradox"– in the *Second European Report* (EC, 1997), as well as Chapter 5 –"Scientific Output and Impact: Europe's Leading Role in World Science"– in the *Third European Report* (EC, 2003a). For a word of caution from the community of official European experts, see the brief comment under the title "From 'European Paradox' to declining competition?" in the EC (2003b) document.

computer science and a number of other disciplines). The first fact on which the 'Paradox' is supposedly based is simply not there. Rather a major EU challenge is how to catch up with the U.S. in terms of scientific excellence." (p. 1455).

In this scenario, this paper contributes to the clarification of the relative position of the EU relative to the U.S. at the turn of the XX century in the light of some novel output indicators of scientific performance.

I. 2. Methods

The standard output indicators used in the literature can be briefly reviewed in two steps.[2] In the first place, a natural performance indicator is the share of publications during a given time period. When there is information on the citations received by these publications, two other indicators are typically added: the share of total citations, and some measure of the citation impact of the average paper.[3] In the second place, the problem with all these indicators is that two distributions that share the same mean might be very differently shaped away from the mean. In our case, it is well known that citation distributions are highly skewed: according to Albarrán *et al.* (2009), for example, about 70% of all articles receive fewer than 30% of all citations while fewer than 10% of them account for more than 40% of all citations. This is surely the reason why the authors in the Leiden group have always completed the average-based indicators used to monitor research groups with the percentage of uncited

[2] For the simultaneous measure of outputs and inputs to the scientific and innovation process, as well as a discussion of productivity indicators, see May (1997, 1998), EU (2003a), King (2004), and Shelton and Holdridge (2004). The latter also includes a review of qualitative methods for the measurement of science and technology consisting of studies of the international stature of research centers in the U.S. and the EU conducted by experts in the corresponding disciplines. For a general discussion of the evolution and shortcomings of science and technology indicators and their use in national policy, see Grupp and Moggee (2004).

[3] There are two types of average-based measures: the impact measures rebased against the world baseline, used *inter alia* in May (1997), Adams (1998), King (2004), EU (2003a), and Shelton and Holdridge (2004), and the relational charts in Glänzel *et al.* (2002) that use information –unavailable in our database– about the journals where each country's articles are published.

papers (Moed *et al.*, 1985, 1988, 1995).[4] More recently, they also include the percentage contribution in the top 5% of most highly cited papers (van Leeuwen *et al.*, 2003). In the same vein, recent contributions to the literature on international comparisons of scientific performance –such as King (2004)– include as an output indicator the country's share in the top 1% of most highly cited papers, information made readily available by TS since 2001 in their *Essential Science Indicators (*http//www.isihighlycited.com).[5]

The distinctive feature of this paper is the computation of the EU and U.S. publication shares at a large number of percentiles of the world citation distribution. The articles published in any given scientific field over a given time period are ordered in increasing number of citations. Then the shares of articles attributed to both geographical areas are computed at every percentile p in the unit interval. When $p = 0.1$, for example, the shares refer to the set of articles after discarding the 10% least cited, or what is the same, to the 90% of more highly cited articles. For a given geographical area, the graph of the publication shares as p increases from 0.1 to 0.2, 0.3, etc., indicates its relative performance as the publications' impact measured by the number of citations increases. The comparisons of such graphs for two geographical areas provide an eloquent picture of their relative situation at many points of the citation distribution, and not only when $p = 0$ and $p = 0.01$ as is the case in recent contributions to this literature.[6]

I. 3. Main Results and Organization of the paper

We are interested in solidly establishing the relative situation of the U.S. and the EU at

[4] The Leiden group also constructs their average-based indicators counting with information about the journals where each country's articles are published. This allows them to compare the research groups' observed mean citation with the expected behavior of the set of journals where the group is known to publish. The ratio of such expected behavior to the behavior of the journals in the entire field constitutes another interesting indicator in this case.
[5] See also Batty (2003) for a study of the pattern of spatial concentration by the highly cited scientists.
[6] The same idea can be found in the study of domestic versus internationally co-authored papers in Glänzel (2000, 2001).

the turn of the century. At the same time, the computation of publication shares at several points of the citation distribution for the smaller fields requires a sizable sample. Therefore, the empirical exercise conducted in this paper refers to the 3,654,675 articles published in 1998-2002 and the more than 47 million citations they receive in 1998-2007. The main results of this paper can be summarized as follows:

(i) The assignment of coauthored articles to a geographical area consisting of several countries must be done with care. If articles attributed to the EU are wrongly computed, then the annual production of the EU becomes upwardly biased by at least 15%.

(ii) The share of articles published in 1998-2002 is equal or larger in the EU than in the U.S. for 15 of the 22 TS scientific fields. However, as soon as the citation impact of these articles is taken into account the picture obtained is very damaging for the EU: the average citation rate in the U.S. is greater than in the EU in every one of the 22 fields.

(iii) In the seven fields, representing 15.5% of the total number of publications, where the share of articles is greater in the U.S. (Molecular Biology and Genetics; Immunology; Neuroscience and Behavior; Psychology and Psychiatry; Economics and Business; Social Sciences, General, and Multidisciplinary), the initial gap between the U.S. and the EU shares widens as we advance towards the more cited articles. Among 14 fields, representing 82.6% of the total, where the share of the total number of publications is equal or larger in the EU than in the U.S., the U.S. surpasses the EU at a low percentile of the world citation distribution in six cases (Biology and Biochemistry; Clinical Medicine; Space Science; Computer Science; Environment and Ecology, and Geoscience); at an intermediate percentile in two cases (Microbiology, and Pharmacology and Toxicology), and at the upper tail of the distribution in six cases (Chemistry; Mathematics; Engineering; Physics; Materials Science, and Plant and Animal Science). Agricultural Sciences, representing 1.9% of the total, is the only field in which the EU dominates the U.S. over the entire distribution.

In brief, it can be concluded that among the most influential articles, in 21 out of 22 scientific fields the dominance of the U.S. over the EU is overwhelming. Thus, although the share of articles published in 1998-2002 in all sciences combined is greater in the EU than in the U.S., the latter overcomes the former when the top 50% of the more highly cited articles is considered.

The rest of the paper is organized in three Sections. Section II presents the data and some yearly statistics for all sciences combined during the period 1998-2007. Section III contains the main empirical findings for each of the 22 scientific fields and all sciences as a whole. Section IV discusses those findings and offers some concluding comments.

II. DATA AND ANNUAL STATISTICS, 1998-2007

II. 1. Data

TS indexed journal articles include research articles, reviews, proceedings papers and research notes. In this paper, only research articles, or simply articles, are studied, so that 390,097 review articles and three notes are disregarded. After the elimination of observations without information about the country or countries where the article was written, or for which other variables were missing, our sample size consists of 8,153,092 articles, or approximately 95% of the number of items in the original database. The citation distribution of the 164,521 articles in Arts and Humanities presented very different characteristics from the remaining TS fields (for example, 83% of all articles received no citations at all). Therefore, the final sample belonging to the 20 fields in the natural sciences and the two social sciences distinguished by TS consists of 7,988,571 articles.

The dataset consists of articles published in a certain year and the citations they receive from that year until 2007, that is, articles published in 1998 and their citations during the 10-year period 1998-2007, articles published in 1999 and their citations in the 9-year period 1999-2007, and so on until articles published in 2007 and their citations during that same year. The total number of citations is 65,042,734.

II. 2. Assignment of Articles to Geographical Areas

Articles are assigned to geographical areas according to the institutional affiliation of their authors as recorded in the TS database on the basis of what had been indicated in the by-line of the publications. Internationally co-authored papers present assignment problems. When the country where each author works is known, one possibility is to assigning each article fractionally to the different countries or geographical areas in proportion to the number of authors working on each of them. For instance, if an article is written by four people, two of them working in the U.S., one working in Denmark, and another one working in Sweden, then 50% of the article would be assigned to the U.S. and 50% to the EU. This rule, followed in EC (1994), is not without problems: the co-authored article in the example would count for less both in the U.S. and in the EU than a similar article with the same number of authors, but all working in either the U.S. or the EU.

At any rate, our database does not have full information about the country where every author works. We only know the countries involved, but not how many authors per country there are. From a U. S. geopolitical point of view, for example, we want to give equal weight to an article written in a U.S. research center than to another co-authored by researchers from a U.S. and a Chinese university, independently in both cases of the number of authors in each area. Thus, as in the classical studies by May (1997) and King (2004), in every internationally co-authored article a full count is credited to each contributing area: articles co-authored by one or more persons affiliated to institutions in either the U.S. and the EU, the U.S. and the rest of the world (RW hereafter), or the EU and the RW, are counted twice, while articles co-authored by persons in the three areas are counted three times. Only articles exclusively authored by one or more persons affiliated to research centers either in the U.S., the EU, or the RW alone, are counted once. In the above example of an article written by four people, two of them working in the U.S., one working in Denmark, and another one working in Sweden, the article is counted twice: once in the U.S. and once in the EU. The total number of articles in such extended count is 9,151,912, or 14.6% more than the standard count in which all articles are counted once. Similarly, the total number of citations in the extended sample is approximately 20% greater than the one in the standard dataset.

Note that, in the presence of geographical areas consisting of several countries, this is the best we can do with the available information. Alternatively, articles in the EU, for example, could be assigned in two steps: first to individual European countries and then to the European aggregate. In the above example, the article would be assigned once to the U.S. and also once to Denmark and Sweden in the first step, and therefore twice to the EU in the second step. Of course, this procedure would artificially blow up the European share to the extent of intra-European cooperation among individual European countries. As illustrated in Figure 1, the amount of the bias could be important.

Figure 1 around here

Figure 1 provides the ratio of publications and citations for all sciences in the comparator area (the EU) to the U.S. in 1998-2007 under the two alternatives, that is, by counting only the articles published in at least one of the EU member countries, or by taking the European publications equal to the sum of the articles published in each member country. It is seen that this second alternative exaggerates the importance of the EU in both publications and citations received. The upward bias in publications starts at 15% and increases with time. These trends indicate that, possibly in response to increased incentives from the European Commission and national sources in favor of intra-European scientific cooperation, co-authorship among researchers working in different European countries clearly increases during 1998-2007. This is an important measurement issue that might be affecting the overtly optimistic view in some quarters about the improved scientific performance in the EU in recent times.[7]

II. 3. Annual Comparisons Among the U.S., the EU, and the RW

Figure 2 illustrates the ratio of publications and citations for all sciences in the comparator area (the EU or the RW) to the U.S. under the correct alternative in 1998-2007. Two comments are in order. First, as has been observed by other authors the EU publishes more scientific articles than the U.S. and is not far behind in total citations. Second, perhaps the more remarkable fact is the rapid growth experienced by the RW, whose articles represented 41% of all those published in the world in 1998 and 52% in 2007; on the other hand, the RW's volume of citations becomes larger than that of the EU from 2004 onwards. The study of these phenomena, however, is beyond the scope of this paper. Our fundamental concern is how the U.S. and the EU shares stand when we compute them at different percentiles of the world citation distribution. The answer to this question for each of the 22 scientific fields and for all of them combined is provided in the next Section.

Figure 2 around here

III. AVERAGE CITATION PER ARTICLE AND PUBLICATION SHARES AT DIFFERENT PERCENTILES OF THE WORLD CITATION DISTRIBUTION, 1998-2002

III. 1. Description of the Sample

As indicated in the Introduction, there are two reasons why we need a large sample.

[7] This could be the case, *inter alia*, of the important contribution by King, 2004, whose Figure 1 (p. 311) states: *"The EU15 total contains some duplication because of papers jointly authored between countries in the EU group."*

Firstly, this paper aims to obtain empirical conclusions at the level of 22 TS fields, but seven of them represent each less than 2% of the total, and another six between 2% and 3%. Thus, thirteen fields might be too small if we were to take only articles published in a single year. Secondly, we want to establish some stylized facts about the relative scientific performance of the U.S. and the EU at the turn of the XX century, when the Lisbon declaration by the European Council took place. This should serve as a benchmark for future comparisons in 20 or 50 years time. Consequently, the remaining part of the paper essentially focuses on the sample of 3,654,675 articles published in 1998-2002 and the 47,239,360 citations they receive in 1998-2007, that is, the maximum citation volume existing in our database. The 20 fields in the natural sciences are organized in three large aggregates: Life Sciences, Physical Sciences, and Other Natural Sciences. The last two represent, approximately, 28.5% and 25.6% of the total, while the Life Sciences represent about 40.7%. The remaining 5.2% correspond to the two Social Sciences.

Table 1 presents information for this sample on two topics: the comparison between the number of original and extended articles in every field, and the percentage distribution by field of the extended number of papers assigned to the U.S. and the EU –the two areas that will be compared in the sequel. It is observed that, on average, in 1998-2002 the extended number of articles assigned to the three geographical areas represent 13.6% more than the original ones. Not surprisingly, the degree of international co-authorship is largest in Space Sciences where the extended count is 38.6% larger than the original one. In Mathematics, Microbiology, Molecular Biology and Genetics, Physics and Geoscience the extended number of articles is between 18% and 22% greater, while in the Social Sciences, Psychiatry and Psychology, Environmental Science and Ecology, and the Multidisciplinary field international co-authorship is relatively less important and the extended number of articles is only between 5%

and 9% greater than the original ones.[8]

Table 1 around here

On the other hand, the percentage distributions by field in the U.S. and the EU are rather close to each other, although the life sciences and the social sciences are slightly more important in the U.S., while the physical sciences are more important in the EU. Relative to a situation in which the two areas were to be heavily specialized in different fields, this correlation makes unnecessary any *a priori* differentiation among fields.

III. 2. Standard Output Indicators

Let x_i be the actual number of articles in field i published in 1998-2002, and denote by y_{ij} the articles in that field assigned to geographical area j = U.S., EU, RW. We will refer to the ratio y_{ij}/x_i as the share of articles in field i written in area j. Denote by $y_i = \Sigma_j y_{ij}$ the extended number of articles in field i. Naturally, $y_i > x_i$, so that in what follows it should be recalled that the sum of the ratios y_{ij}/x_i add up to more than one.[9] Similarly, let c_i be the actual number of citations received by the x_i articles in field i, and denote by d_{ij} the number of citations received by the y_{ij} articles in area j. We will refer to the ratio d_{ij}/c_i as the share of total citations in field i received by area j. Again, since $d_i = \Sigma_j d_{ij} > c_i$, the sum of the ratios d_{ij}/c_i add up to more than one. Finally, the ratio

$$\mu_{ij} = [d_{ij}/c_i]/[y_{ij}/x_i] = [d_{ij}/y_{ij}]/[c_i/x_i]$$

is the average number of citations per article in field i and area j, normalized by the actual mean citation rate in the world as a whole. We will refer to μ_{ij} as the normalized mean citation

---

[8] Citations to the extended articles, for which information by field is available on request, represent 19.7% more than citations to the original articles

[9] For example, as pointed out before, for all sciences combined in 1998-2002 the sum will be equal to 1.136, indicating that the extended number of articles is 13.6% greater than the actual number of them.

rate in field i and area j. Of course, a value of $\mu_{ij} = 1.2$ (or $\mu_{ij} = 0.95$), for example, means that the average citation rate in field i is 20% higher (or 5% lower) in area j than in the world.

Table 2 presents the following standard output indicators of scientific performance for the U.S. and the EU in every field: (i) the share of articles published in 1998-2002, y_{ij}/x_i, (ii) the share of total citations received by these articles in 1998-2007, d_{ij}/c_i, and (iii) the normalized mean citation rate per article, μ_{ij}. Scientific fields are classified in two groups: group I includes fields where the share of articles in the U.S. is greater than the share in the EU, while group II includes fields where the opposite is the case.

Table 2 around here

As we know from Figure 2 and as can be observed in the last row and columns 1 and 2 in Table 2, the share of articles in all fields combined in 1998-2002 is greater in the EU than in the U.S. But this hides differences across fields that it is important to highlight. To begin with, among group I fields there are twice as many Social Science articles in the U.S. as there are in the EU. Taking into account that these disciplines are largely devoted to nationally defined issues, and that TS covers journals in English but not so well at all journals in other languages, this large quantitative superiority of the U.S. over the EU should come as no surprise. Something similar can be said about Psychology and Psychiatry. The remaining four cases in group I are Neuroscience and Behavior, Molecular Biology and Genetics, and Immunology among the life sciences, plus the Multidisciplinary field. From the 8th field (Agricultural Sciences) to the 14th (Microbiology) in group II, the EU share is at least nine percentage points greater than the U.S. share. From the 15th field (Pharmacology and Toxicology) to the 20th (Engineering) that difference goes down to three to six percentage points. In the last two group II fields (Biology and Biochemistry and Environment and Ecology) the EU and U.S.

shares are practically equal.

The key fact is that as soon as we turn from the sheer production of scientific articles toward the impact they have in terms of total citations received, the relative situation of the two geographical areas is dramatically reversed: for all sciences combined, the share of total citations of U.S. articles in our 1998-2002 dataset is almost seven percentage points greater than the EU share (see the last row and columns 3 and 4 in Table 2). This is partly the result of the clear U.S. superiority in all group I fields already discussed. In the last seven group II fields, where the EU articles share is quite close or equal to the U.S. one, the total citation share is greater in the U.S. Only in the remaining eight, where there are considerably more EU than U.S. publications, the EU citation share is equal to or greater than that of the U.S.. From another point of view, the total citation share is much greater in the U.S. for the Social Sciences and the Life Sciences, about 4 percentage points greater in the EU for the Physical Sciences, and equal for the other natural sciences.

Finally, when we focus on the normalized mean citation rate (MCR hereafter) in columns 5 and 6 in Table 2, the comparison becomes completely one-sided: for the articles published during 1998-2002 the U.S. dominates the EU in every one of the 22 scientific fields. In the first place, except for the Multidisciplinary field, the MCR in all group I fields in the EU is just equal to or smaller than that of the world as a whole. The distance with the U.S. in this group, however, is of 19-69 percentage points. In the second place, within group II, the MCR for the EU is considerably greater than that of the world in only eight fields (Agricultural Sciences, Plant and Animal Science, Physics, Chemistry, Geoscience, Engineering, Mathematics, and Materials Science). Even in these cases the distance with the U.S. goes from seven percentage points in Plant and Animal Sciences to 43 in Chemistry. In the third place, the EU's performance is also particularly dismal, just below or above the world standard, in other 3 group II fields (Computer Science, Clinical Medicine, and Biology and Biochemistry). Looking at group II as a whole, where the EU's share of articles is greater than or equal to that of the U.S., it can be concluded that the EU publishes too many poorly cited articles. As a result, the MCR for all sciences combined in the EU is only 8% above the world one and 33 percentage points below the U.S. one.

The total dominance by the U.S. in MCRs is lost in the influential paper by King (2004). As indicated in the Introduction, this author states that "*the EU now matches the United States in the physical sciences, engineering and mathematics, although still lags in the life sciences*". But this statement refers to the share of total citations (Figure 4, p. 315), a fact essentially confirmed in columns (3) and (4) of our Table 2. However, once the number of articles is also taken into account, all MCRs become greater in the U.S. The reason is that, as can be seen in columns (1) and (2) in Table 2, the number of publications in these fields is greater in the EU.

The average statistics just reviewed are very informative. However, as indicated in the Introduction, given how skewed citation distributions generally are it is important to make comparisons at different impact levels and, most specially, among the highly cited articles at the upper tail of the citation distribution. This is what is done in the next sub-section.

III. 3. Publication Shares at Different Percentiles of the Citation Distribution

In every field, let us order the 3,654,675 original articles published in 1998-2002 by the number of citations they receive. Recall that articles published in each of those years receive citations over a different time period. Therefore, in order not to discriminate in favor of earlier published articles that receive citations over a longer time period, we first partition the articles published in each of these five years in percentiles, and then construct each percentile for

1998-2002 by adding up the corresponding articles published in each year.[10] Finally, in each percentile compute the share of (extended) articles with at least one author working in a research institution in the U.S., the EU or the RW with respect to the total actual number of articles in each percentile.[11] The U.S. and EU publication shares at different percentiles of the citation distribution for the 22 fields and for all fields combined are presented in Figures 3 to 6.[12]

Figures 3 to 6 around here

Figure 3 includes the seven group I fields. As can be observed, in every case the initial gap between the U.S. and the EU shares widens up as p increases. Figure 4 includes eight group II fields in which the initial publication share is greater in the EU than in the U.S.. The distinctive feature here is that the U.S. surpasses the EU rather early in six cases (Biology and Biochemistry at $p = 0.10$; Space Science at $p = 0.25$; Environment and Ecology at $p = 0.35$; Clinical Medicine at $p = 0.36$; Computer Science at $p = 0.46$); or at intermediate points (Geoscience at $p = 0.55$; Microbiology at $p = 0.70$, and Neuroscience and Behavior at $p = 0.71$). Figure 5 includes six group II fields in which the U.S. surpasses the EU rather late (Chemistry at $p = 0.90$; Mathematics at $p = 0.92$; Engineering at $p = 0.94$; Physics and Material Sciences at $p = 0.95$, and Plant and Animal Sciences at $p = 0.96$). Finally, Figure 6 contains the Agricultural Sciences, the only case in which the EU dominates the U.S. at every percentile, as well as all sciences combined in which the U.S. surpasses the EU at about $p = 0.50$.

Two final points are in order. Firstly, the average results in terms of MCRs and the relative performance illustrated in Figures 3 to 6 are consistent with each other. The eight

[10] This is also the method followed in the construction of the top 1% more highly cited articles in the Web of Science's *Essential Science Indicators*.
[11] As before, the sum of such shares at every percentile will not add up to one.
[12] As a matter of fact, only the following 21 percentiles are actually computed in each case: 0, 10, 15,…, 90, 95, 98, and 99.

fields for which the European MCR is well above the world level appear in the later Figures, namely, Geoscience in Figure 4, six fields in Figure 5 and, of course, Agricultural Sciences in Figure 6. On the other hand, group I fields where most European MCRs are below world levels is when the U.S. dominance is truly overwhelming in Figure 3. Secondly, the U.S. curves tend to have a positive slope and, when the upper tail is reached at $p = 0.90$, they all clearly rise without exception. However, in about 10 fields the EU share remains relatively flat or slightly increases, while in the remaining 12 decreases at that crucial stage.

IV. DISCUSSION AND CONCLUSIONS

As documented in the Introduction, different people and institutions held rather different views about the relative scientific performance of the EU and the U.S. at the turn of the XXI century. As a contribution to the settlement of this issue, together with standard output indicators such as the MCR, this paper has compared the U.S and EU publication shares at every percentile of the world citation distribution. The data used is a sample of 3.6 million articles published in 1998-2002 in 22 scientific fields and the more than 47 million citations they receive in 1998-2007. The idea has been to build a benchmark case to be evaluated with similar techniques in 20 or 50 years time.

The facts of the matter can be summarized in three points. Firstly, it has been shown that an incorrect assignment of articles to geographical areas including several countries, such as the EU, can lead to an upward bias in the total number of articles of at least 15%. After applying the best possible assignment methods, the EU share of total publications in all sciences in 1998-2002 is about 4% greater than that of the U.S.. Secondly, as soon as these articles' impact, measured by the citations they receive, is taken into account, the overall picture is reversed: the EU MCR for all sciences combined is only 8% above the world rate, but 33 percentage points below that of the U.S.. Moreover, the U.S. publication share becomes greater than the EU's for the top 50% of the more highly cited articles. Thirdly, there are of course differences across fields. In particular, the EU performs well above the world average in eight fields: Agricultural Sciences, as well as Plant and Animal Science; Physics, Chemistry, and Mathematics; Geoscience, Materials Science, and Engineering. However, the European MCR is considerably greater than the U.S. one in all of these favorable cases, and the EU publication share in all these fields is surpassed by the U.S.' for all percentiles beyond the top 45% or the top 4% of the more highly cited articles, depending on the case. On the other hand, relative to the world and the U.S. the EU performance is particularly poor in the following cases: the Social Sciences –nor surprisingly, given the nature of the data and of the fields themselves– Computer Science, and all life sciences independently of whether the total publication share is greater in the U.S or the EU.

The overall conclusion is inescapable. Independently of sectoral details just discussed, according to our large 1998-2002 dataset acquired from TS one must fully side with Dossi *et al.*'s (2006) diagnosis when they argue that *"one observes a structural lag in top-level science vis-a-vis the US"*.

This is not the place for the formulation of policy recommendations. But we may get closer to that aim after carrying a number of extensions. In the first place, one should recall that in all fields the articles assigned to any particular geographical area are of two types: those fully written in institutions within that area, and those co-authored with someone working somewhere else in the world. The separate investigation of the impact of these two types of articles along the lines advocated in this paper would help us to learn about the existence and importance of the potential benefits from co-authorship in the different disciplines.

In the second place, the present analysis should be extended in rather obvious directions: towards specific sub-fields, countries, and even individual research centers. In the EU case, for example, these extensions are advisable both when things do not go particularly well, as in the field of Clinical Medicine that represents 23% of the total, or in fields where the European performance has been shown to be better.

In the third place, the present study of the publication share at every percentile of the world distribution has registered what we may call the *incidence* of the low- and high-impact aspects of a geographical area citation distribution. But we may also focus on two other interesting features. Firstly, the *intensity* of the low- and high-impact aspects of citation distributions. Consider a critical citation level (CCL) fixed, for example, at the 80^{th} percentile of the world citation distribution. Then low- and high-impact gaps can be naturally defined: the first as the difference between the CCL and the citations received, and the second as the difference between the citations received and the CCL. The aggregation of such gaps into a scalar according to different procedures would yield useful measures of the intensity of the low- and high-impact phenomena. Secondly, one may also worry about citation *inequality* among low- and high-impact articles. This program, inspired in the analysis of economic poverty in income distributions, is carried on in Albarrán *et al.* (2009b, c).

In the fourth place, so far we have insisted on looking at the entire citation distribution in each field. However, there is no doubt that the most relevant basic and applied research is generally found at the very top of citation distributions. Consequently, robust measures of scientific excellence, such as the *h*-index and its many variants may be particularly appropriate. Contrary to the output indicators used in this paper, these measures of excellence are size dependent in the sense that that they cannot but increase with the size of the set of articles under evaluation. Thus, size normalization, such as the procedures pioneered in Katz (1999, 2000) and Molinari and Molinari (2008a, 2008b), are called for in this case.

Among others, these extensions may help close the gap toward specific policy recommendations. However, as a final thought we may reflect on the policy implications of the following phenomenon already stressed in the text: in every field without exception the U.S. publication share tends to increase from the beginning, and specially when the top 10% of more highly cited articles is reached, giving rise to a clearly convex curve in Figures 3 to 6; however, little of the sort is exhibited in the EU case –a pattern, it should be emphasized, that could hardly be ascertained with the sole help of average-based indicators. It would appear that the behavior of the U.S. scientific community at the upper tail of the citation distribution in all fields calls for some systematic explanation in terms of the institutions that are known to work particularly well in that country: the strong competition among the top public and private universities and research centers in search for excellence and world wide recognition, and the incentives of all sorts used to attract the best scientists from all over the world and to extract the most from them once they become full-time players in their institutions.

REFERENCES

Albarrán, P., J. Crespo, I. Ortuño and J. Ruiz-Castillo (2009a), "References Made and Citations Received By Scientific Articles", mimeo.

Albarrán, P., I. Ortuño and J. Ruiz-Castillo (2009b), "The Measurement of Low- and High-impact in Citation Distributions: Technical Results", Working Paper 09-57, Economics Series 35, Universidad Carlos III.

Albarrán, P., I. Ortuño and J. Ruiz-Castillo (2009c), "Low- and High-impact Citation Measures: Empirical Applications", Working Paper 09-57, Economics Series 35, Universidad Carlos III.

Adams, J. (1998), "Benchmarking International Research", *Nature*, **396**: 615-618.

Batty, M. (2003), "The Geography of Scientific Citation", *Environment and Planning A*, **35**: 761-770.

Dosi, G, P. Llerena, and M. Sylos Labini (2006), "Science-Technology-Industry Links and the 'European Paradox': Some Notes on the Dynamics of Scientific and Technological Research in Europe", *Research Policy*, **35**:1450–1464.

EC (1994), *First European Report on Science and Technology Indicators*, Directorate-General XII, Science, Research, and Development. Luxembourg: Office for Official Publications of the European Community.

EC (1997), *Second European Report on Science and Technology Indicators*, Directorate-General XII, Science, Research, and Development. Luxembourg: Office for Official Publications of the European Community.

EC (2003a), *Third European Report on Science and Technology Indicators*, Directorate-General for Research. Luxembourg: Office for Official Publications of the European Community, http://www.cordis.lu/rtd2002/indicators/home.htlm

EC (2003b), "From 'European Paradox' to Declining Competitiveness?", Snapshots, 4, in *Key Figures 2003/2004*, Directorate-General for Research. Luxembourg: Office for Official Publications of the European Community, http://cordis.europa.eu/indicators/publications.htm

Glänzel, W. (2000), "Science in Scandinavia: A Bibliometric Approach", *Scientometrics*, **48**: 121-150.

Glänzel, W. (2001), "National Characteristics in International Scientific Co-authorship Relations", *Scientometrics*, **51**: 69-115.

Glänzel, W., A. Shubert, and T. Braun (2002), "A Relational Charting Approach to the World of basic Research in Twelve Science Fields at the End of the Second Millennium", *Scientometrics*, **55**: 335-348.

Grupp, H. and M. E. Mogee (2004), "Indicators for National Science and Technology Policy: How Robust are Composite Indicators?", *Research Policy*, **33**: 1373-1384.

Katz, J. Sylvan (1999), "The self-similar science system", *Research Policy*, **28**: 501-517.

Katz, J. Sylvan (2000), "Scale-independent indicators and research evaluation", *Science and Public Policy*, **27**: 23-36.

King, D. (2004), "The Scientific Impact of Nations", *Nature*, **430**: 311-316.

May, R. (1997), "The Scientific Wealth of Nations", *Science*, **275**: 793-796.

May, R. (1998), "The Scientific Investments of Nations", *Science*, **281**: 879-880.

Moed, H. F., W. J. Burger, J. G. Frankfort, and A. F. J. van Raan (1985), "The Use of Bibliometric Data for the Measurement of University Research Performance", *Research Policy*, **14**: 131-149.

Moed, H. F., and A. F. J. van Raan (1988), "Indicators of Research Performance", in A. F. J. van Raan (ed), *Handbook of Quantitative Studies of Science and Technology*, North Holland: 177-192.

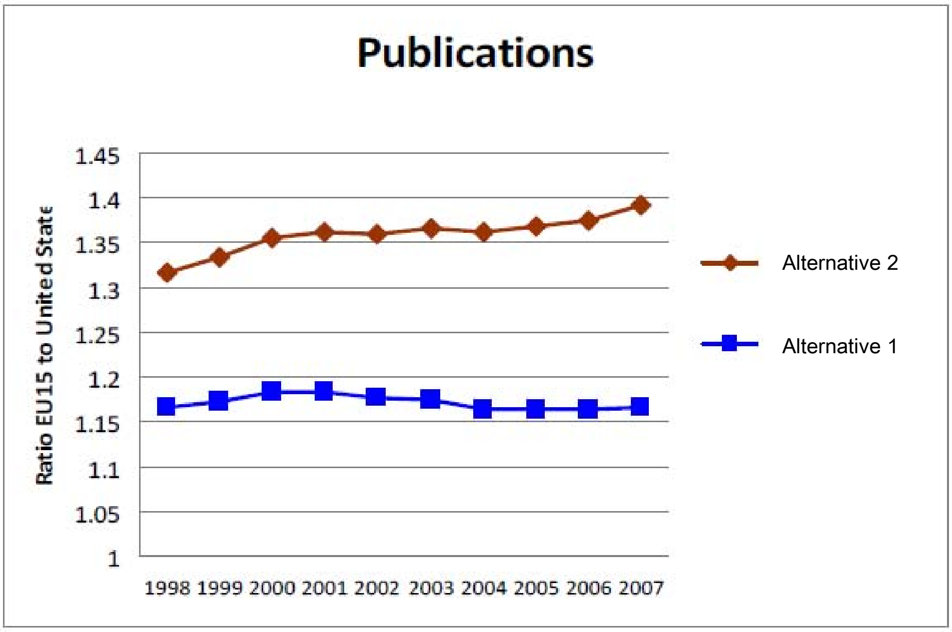
Moed, H. F., R. E. De Bruin, and Th. N. van Leeuwen (1995), "New Bibliometrics Tools for the Assessment of national Research Performance: Database Description, Overview of Indicators, and First Applications", *Scientometrics*, **133**: 381-422.

Molinari, A. and J.-F. Molinari (2008a), "A New Methodology for ranking Scientific Institutions", *Scientometrics*, **75**: 163-174.

Molinari, A. and J.-F. Molinari (2008b), "Mathematical Aspects of a New Criterion for Ranking Scientific Institutios Based on the *h*-index", *Scientometrics*, **75**: 339-356.

Shelton, R. and G. Holdridge (2004), "The US-EU Race for Leadership of Science and Technology: Qualitative and Quantitative Indicators", *Scientometrics*, **60**: 353-363.

Van Leeuwen, T., M. Visser, H. Moed, T. Nederhof, and A, van Raan (2003), "The Holy Grail of Science Policy: Exploring and Combining Bibliometric Tools In Search of Scientific Excellence", *Scientometrics*, **57**: 257-280.



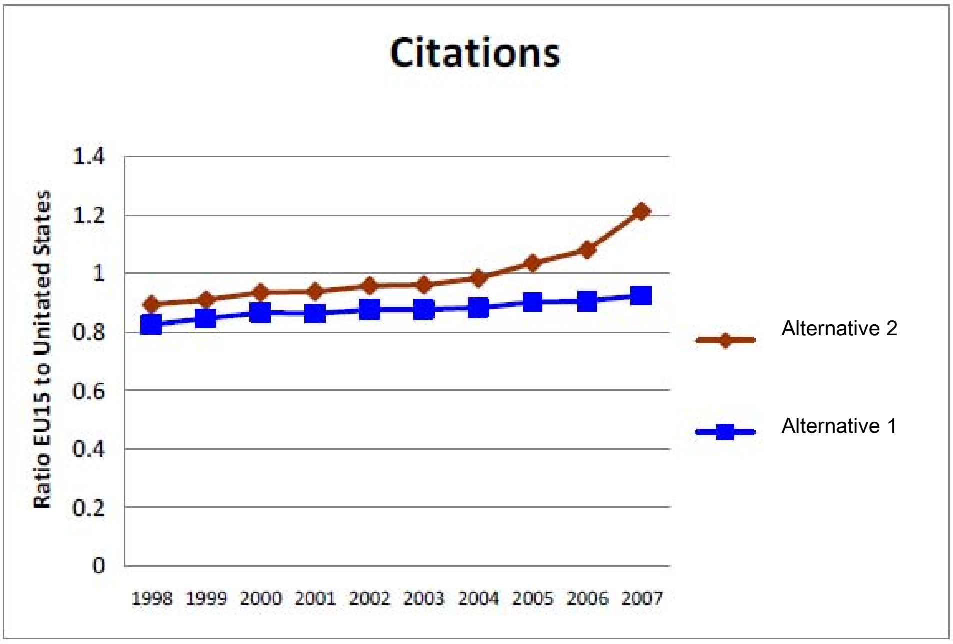


Figure 1. Two Assignment Procedures for the EU, 1998- 2007

Alternative 1. Articles in all sciences with at least one author from the EU. Alternative 2. EU in two steps: first, articles in all sciences with at least one author in any of the 15 member countries; second, EU as the sum of articles in each of the member countries.

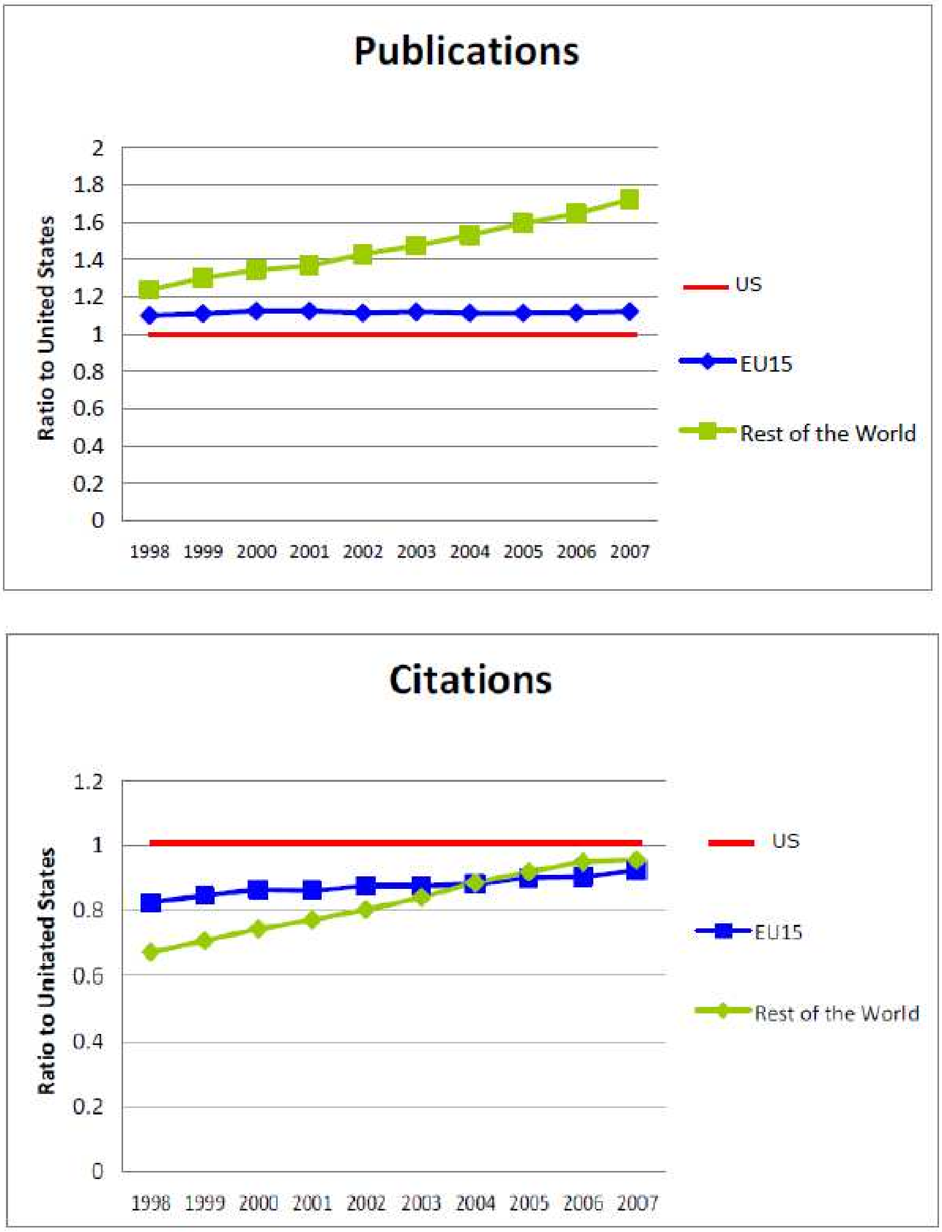


Figure 2. Ratio of Publications and Citations in All Sciences in the EU and the Rest of the World in Relation to the U.S., 1998-2007

Table 1. Extended versus Original Number of Articles, and Extended Distributions by Field in the U.S. and the EU, 1998-2002

FIELDS	Original Articles (1)	Extended Articles (2)	(3)= (1 – 2)/(1) In %	Number of Extended Articles EU	%	USA	%
LIFE SCIENCES	**1,487,856**	**1,672,142**	**12.4**	**583,390**	**43.5**	**565,370**	**47.0**
1. Clinical Medicine	779,597	859,010	10.2	31,086	23.6	279,776	23.3
2. Biology & Genetics	226,851	261,303	15.2	83,941	6.3	82,560	6.9
3. Neuroscience & Behavior	115,199	131,801	14.1	45,361	3.4	47,414	3.9
4. Molecular Biology	101,212	120,994	19.5	39,716	2.9	45,541	3.8
5. Psychiatry & Psychology	90,619	98,446	8.6	27,738	2.1	49,490	4.1
6. Pharmacology & Toxicology	63,103	70,010	10.9	22,676	1.7	18,479	1.5
7. Microbiology	60,250	69,642	15.6	25,545	1.9	19,967	1.7
8. Immunology	51,025	60,936	19.4	21,327	1.6	22,143	1.8
PHYSICAL SCIENCES	**1,040,097**	**1,216,374**	**16.9**	**389,102**	**29.0**	**262,488**	**21.8**
9. Chemistry	450,245	500,897	11.2	155,842	11.6	93,631	7.8
10. Physics	373,248	453,087	21.4	143,442	10.7	92,397	7.7
11. Computer Science	71,834	81,062	12.8	27,460	2.0	25,247	2.1
12. Mathematics	95,554	113,118	18.4	38,059	2.8	29,110	2.4
13. Space Science	49,216	68,210	38.6	24,299	1.8	22,103	1.8
OTHER NATURAL SCIENCES	**936,051**	**1,056,775**	**12.9**	**316,166**	**23.5**	**267,987**	**22.3**
14. Engineering	287,750	320,343	11.3	91,836	6.8	85,565	7.1
15. Plant & Animal Science	215,056	243,455	13.2	73,771	5.5	64,330	5.3
16. Materials Science	162,143	180,150	11.1	52,647	3.9	31,211	2.6
17. Geoscience	96,772	117,982	21.9	37,550	2.8	33,350	2.8
18. Environment & Ecology	88,567	100,939	13.9	31,684	2.4	32,118	2.7
19. Agricultural Sciences	67,110	73,432	9.4	24,824	1.8	16,557	1.4
20. Multidisciplinary	18,653	20,474	9.8	3,854	0.3	4,856	0.4
SOCIAL SCIENCES	**190,671**	**205,286**	**7.7**	**52,702**	**3.9**	**106,569**	**8.9**
21. Social Sciences, General	138,976	146,705	5.6	35,145	2.6	78.916	6.6
22. Economics & Business	51,695	58,581	13.3	17,557	1.3	27.653	2.3
ALL SCIENCES	**3,654,675**	**4,150,577**	**13.6**	**1,341,360**	**100.0**	**1,202,414**	**100.0**

Table 2. Standard Output Indicators of Scientific Performance in the U.S. and in the EU, 1998-2002

SCIENTIFIC FIELDS	Share of Articles		Share of Total Citations		Normalized Average Citation	
	U.S. (1)	EU (2)	U.S. (3)	EU (4)	U.S. (5) = (3)/(1)	EU (6) = (4)/(2)
GROUP I						
1. Social Sciences, General	56.8	25.3	64.9	24.0	1.14	0.95
2. Economics & Business	53.5	34.0	70.8	28.4	1.32	0.84
3. Psychiatry & Psychology	54.6	30.6	63.4	29.2	1.16	0.95
4. Neuroscience & Behavior	41.2	39.4	53.3	38.7	1.29	0.98
5. Molecular Biology and Genetic	45.0	39.2	59.3	39.4	1.32	1.00
6. Immunology	43.4	41.8	53.3	40.3	1.23	0.97
7. Multidisciplinary	26.0	20.7	53.9	28.4	2.07	1.38
GROUP II						
8. Agricultural Sciences	24.7	37.0	31.5	44.1	1.28	1.19
9. Materials Science	19.2	32.5	28.8	35.3	1.49	1.09
10. Plant & Animal Science	29.9	34.3	37.0	40.0	1.24	1.17
11. Chemistry	20.8	34.6	32.3	38.8	1.55	1.12
12. Physics	24.7	38.4	38.7	44.5	1.56	1.16
13. Mathematics	30.5	39.8	41.4	44.0	1.36	1.10
14. Microbiology	33.1	42.4	44.3	44.8	1.34	1.06
15. Pharmacology & Toxicology	29.3	35.9	38.2	38.3	1.30	1.07
16. Clinical Medicine	35.9	40.7	48.9	40.6	1.36	1.00
17. Space Science	44.9	49.4	63.7	53.3	1.42	1.08
18. Computer Science	35.1	38.3	51.6	36.2	1.47	0.95
19. Geoscience	34.5	38.8	47.8	43.2	1.39	1.11
20. Engineering	29.7	31.9	37.7	35.4	1.27	1.11
21. Biology and Biochemistry	36.3	37.0	49.6	37.4	1.36	1.01
22. Environment & Ecology	36.3	35.8	43.3	38.6	1.20	1.08
SOCIAL SCIENCES	55.9	27.6	66.9	25.5	1.20	0.92
PHYSICAL SCIENCES	25.2	37.4	37.9	42.0	1.50	1.12
OTHER NATURAL SCIENCES	28.6	33.8	38.2	38.7	1.34	1.15
LIFE SCIENCES	38.0	39.2	51.0	39.3	1.34	1.00
ALL SCIENCES	**32.9**	**36.7**	**46.3**	**39.5**	**1.41**	**1.08**

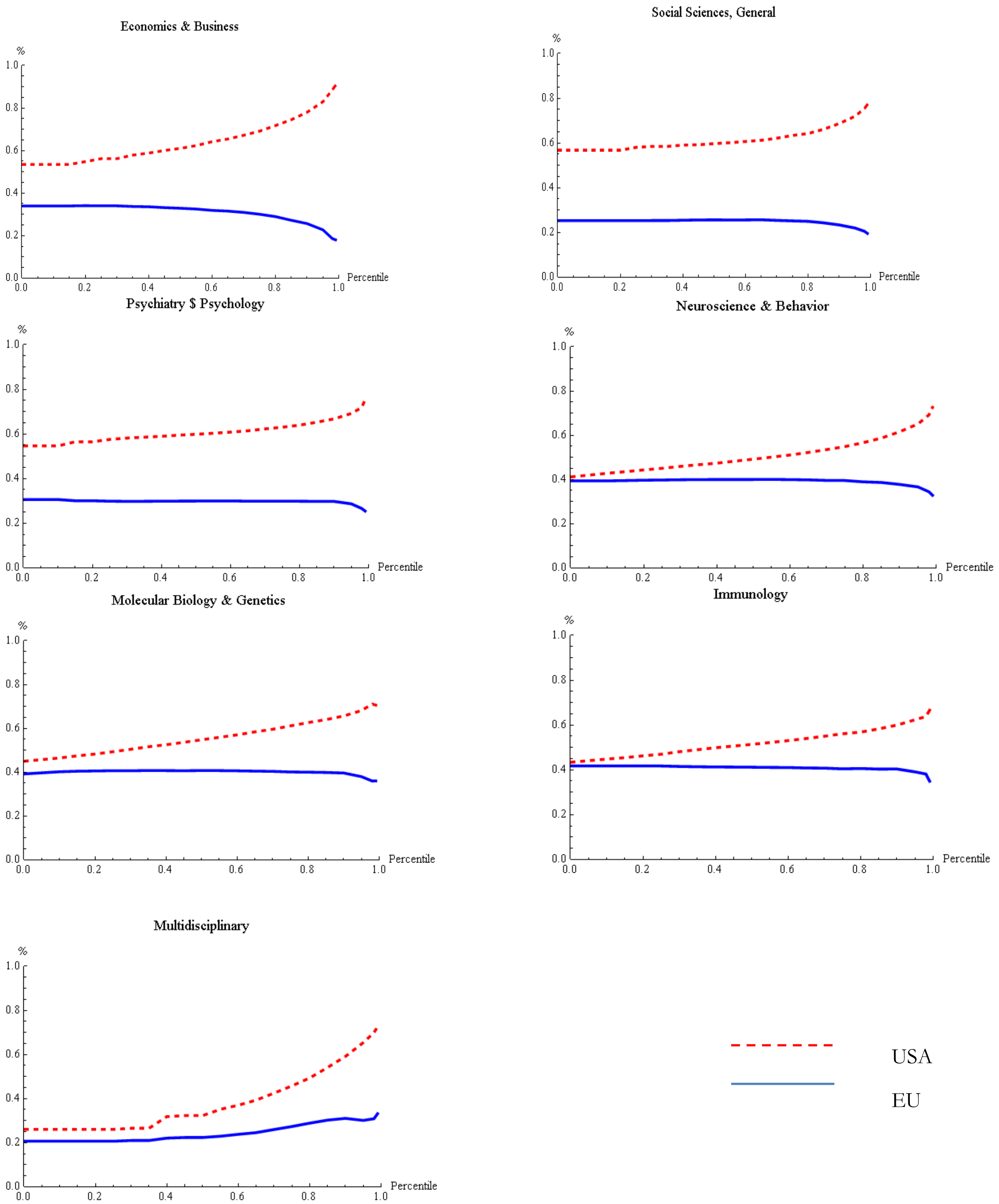


Figure 3. Very Clear U.S. Superiority Over the EU

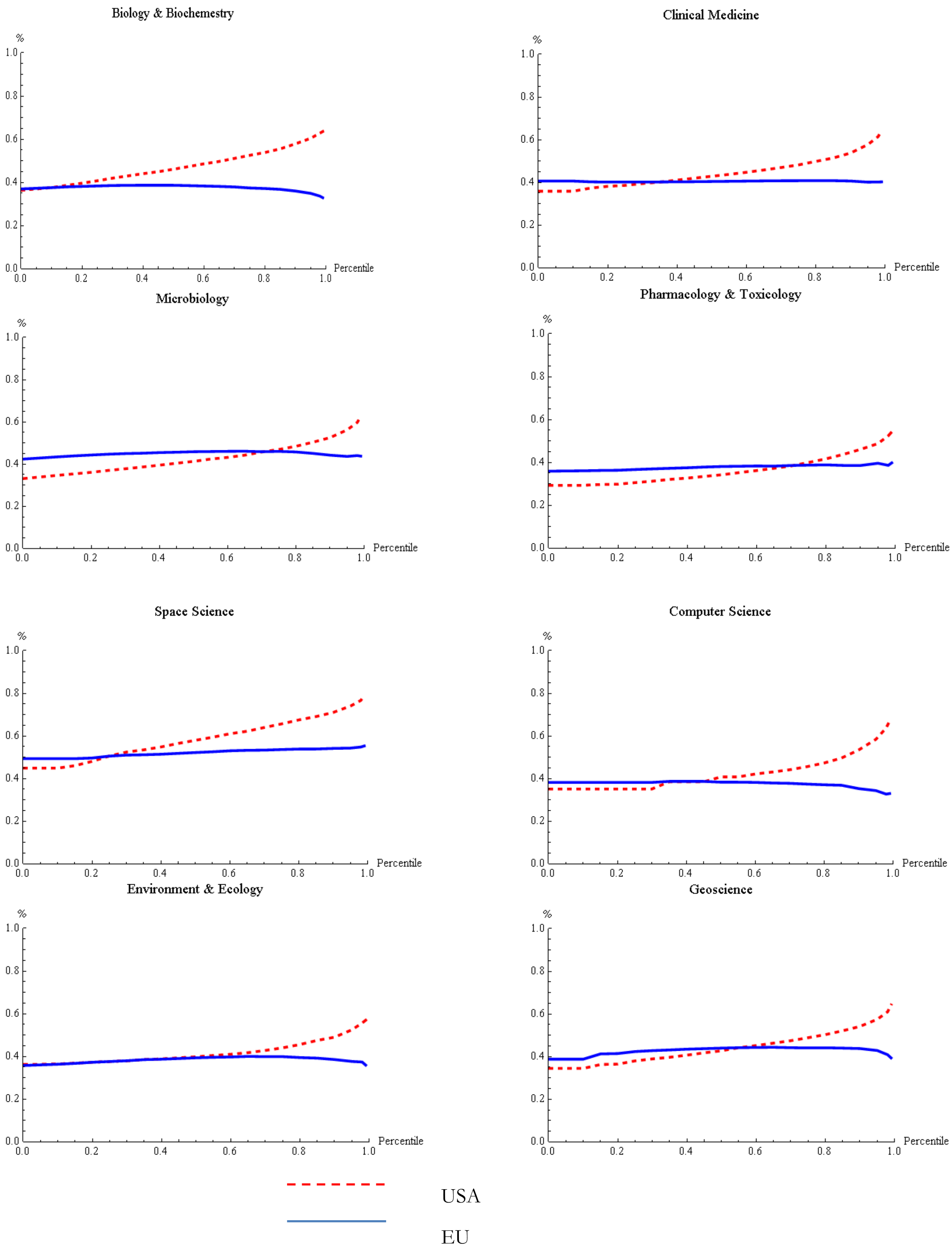


Figure 4. Relatively Clear U.S. Superiority Over the EU

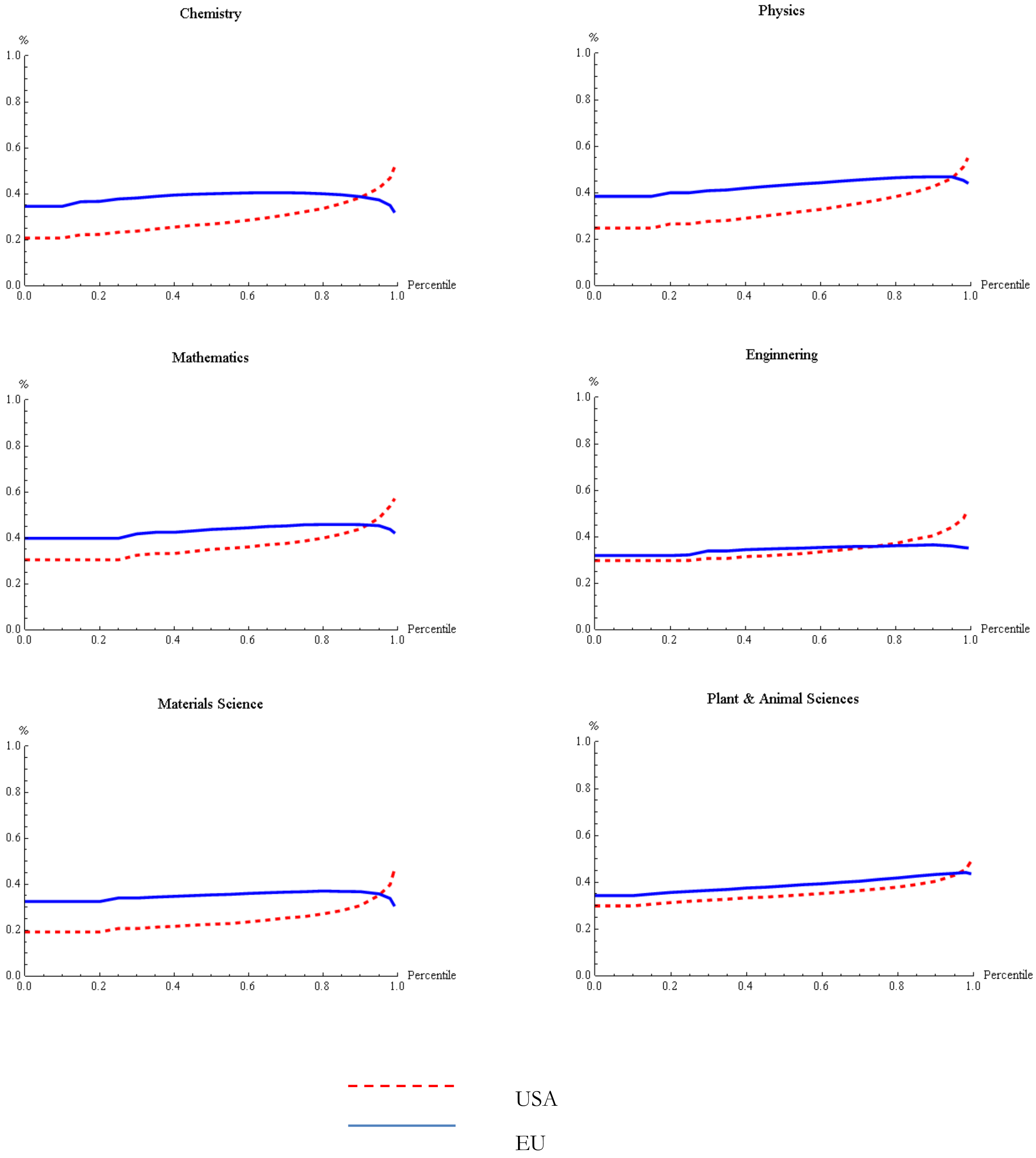


Figure 5. Slight U.S. Superiority Over the EU

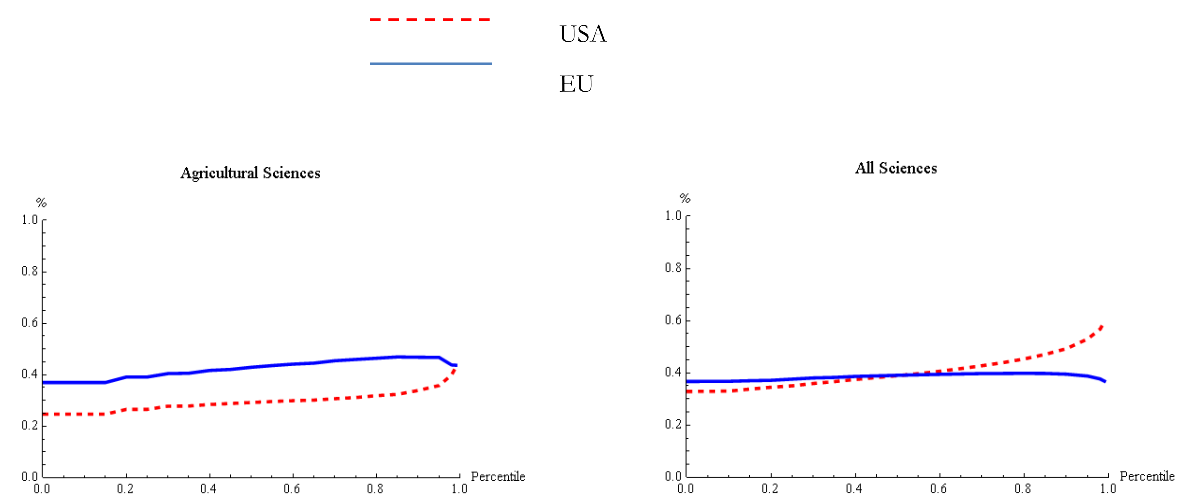


Figure 6. The Only Field with EU Superiority Over the U.S., and the Case for All Sciences Combined